L'"infaillibilité" de l'introspection : autour de Dennett et Wittgenstein

Autor(en): **Bouveresse, Jacques**

Objekttyp: **Article**

Zeitschrift: **Revue de Théologie et de Philosophie**

Band (Jahr): **40 (1990)**

Heft 2

PDF erstellt am: **13.07.2024**

Persistenter Link: https://doi.org/10.5169/seals-381410

REVUE DE THÉOLOGIE ET DE PHILOSOPHIE, 122 (1990), P. 217-233

L' «INFAILLIBILITÉ» DE L'INTROSPECTION

Autour de Dennett et Wittgenstein

JACQUES BOUVERESSE

La question de la conscience: le «petit homme» dans le cerveau

Dans un roman de Patricia Highsmith, le héros, Vic, que sa femme, Melinda, soupçonne, sur la foi de ce qu'elle a entendu dire par un psychiatre, d'être atteint de schizophrénie, se trouve dans la situation d'avoir à expliquer à sa fille de six ans, Trixie, ce que signifient le mot «schizophrénie» et un certain nombre d'autres termes du vocabulaire psychologique. La définition du mot «conscience» soulève une difficulté particulière, qui se résout de la façon suivante: «Vic avait préparé le dîner dont Melinda ne mangea pas grand-chose et elle titubait sous l'effet de l'ivresse à neuf heures, l'heure pour Trixie d'aller au lit. A ce moment-là Vic avait défini plusieurs autres termes psychologiques. Il était difficile d'expliquer à Trixie ce qu'est la conscience, mais il lui dit que lorsque les gens ont eu trop à boire et qu'ils tombent endormis sur le sofa ils souffrent de la perte de celle-ci»[1].

La conscience est probablement, comme le temps selon saint Augustin, l'une de ces choses dont on sait parfaitement ce qu'elles sont aussi longtemps qu'on ne nous demande pas de l'expliquer, et dont on ne le sait plus, dès lors que l'on doit essayer de donner une explication. Le père choisit ici une méthode qui aurait certainement plu à Wittgenstein et qui consiste à expliquer que la conscience est tout simplement ce que l'on perd, lorsque, pour une raison ou pour une autre, on est affecté d'une perte de conscience. Mais un philosophe de type traditionnel objecterait certainement qu'on n'a pas expliqué la conscience si l'on se contente d'indiquer dans quels cas les gens cessent d'être ce qu'ils sont supposés être pendant tout le reste du temps, à savoir «conscients».

A première vue, la perte de conscience est la perte d'un certain type de connaissance qui a la particularité d'être immédiat et infaillible. Le lieu commun en vertu duquel nous disposons d'un accès absolument direct et totalement sûr à nos propres contenus de conscience et qui fait que nous ne pouvons pas commettre d'erreur à leur sujet a été exprimé et expliqué de

[1] *Deep Water*, Penguin Books, 1974, p. 161.

différentes façons par les philosophes. Le problème n'a été résolu généralement que par l'introduction d'un personnage que Daniel Dennett appelle l' «introspecteur infaillible», qui semble être devenu mythique pour la plupart des théories de la conscience contemporaines et qu'elles s'efforcent d'expulser du scénario, le plus souvent sans se rendre compte qu'une fois dépouillé de son rôle d'introspecteur, le «petit homme dans le cerveau» ne cesse de réapparaître sous des formes différentes. Le mythe du spectateur privilégié, qui rapporte des choses dont il est une sorte de témoin oculaire, par opposition à tous les autres gens, qui ne les connaissent qu'à travers ce qu'il en dit, a été rejeté par la plupart des conceptions actuelles, qui le considèrent comme naïf et confus. Mais, comme le remarque Dennett, il n'est pas du tout certain que les explications qui ont été proposées pour le remplacer soient beaucoup plus satisfaisantes.

Puisque l'image de l'introspecteur infaillible qui observe directement des contenus de conscience est supposée être due en dernière analyse à Descartes, tout le problème est de savoir si les analyses récentes ont réussi à remplacer la conception que l'on peut appeler «cartésienne», au sens large, par quelque chose de plus convaincant. Comme le remarque Dennett: «Les rivaux les plus prometteurs de la conception cartésienne partent tous de l'observation que, puisque tout compte rendu *(report)* référentiel, factuel peut être erroné, nos énonciations introspectives ne doivent pas être des comptes rendus référentiels, factuels. C'est ainsi que Wittgenstein soutient (ou l'on soutient fréquemment qu'il soutient) que l'invulnérabilité à l'erreur des comptes rendus de *douleur* (*pain* reports) est due au fait que 'l'expression de la douleur remplace le fait de pousser des cris et ne le décrit pas' et par conséquent n'est pas du tout un compte rendu, mais s'apparente à d'autres manifestations comportementales comme se tordre ou crier. Ryle adopte une position semblable dans *The concept of mind,* où il dit que les comptes rendus de douleur sont des 'aveux' *(avowals)* et non des assertions. La solution de Miss Anscombe consiste à affirmer que les comptes rendus de douleur et certains autres comptes rendus introspectifs ne constituent pas des cas dans lesquels nous avons une *connaissance* de ce que nous disons, mais dans lesquels simplement nous *pouvons dire* ce que nous disons: «il y a un intérêt à parler de connaissance uniquement là où un contraste existe entre 'il sait' et 'il croit (simplement) qu'il sait' » [2]. Ces conceptions ont toutes en commun la démarche qui consiste à faire des comptes rendus introspectifs le genre de chose auxquelles 'correct' et 'incorrect' ou 'vrai' et 'faux' ne s'appliquent pas, mais elles sont implausibles pour des raisons diverses»[3]. Bien entendu, les gens qui défendent des

[2] *Intention,* B. BLACKWELL, Oxford, 2e éd., 1958, p. 14.

[3] *Content and Consciousness,* Routledge and Kegan Books, 2e éd., Londres, 1986, p. 100.

conceptions de ce genre savent parfaitement qu'une déclaration comme «J'ai mal» peut en réalité bel et bien être fausse, par exemple si son auteur veut faire croire qu'il a mal, alors que ce n'est pas le cas. Ce qu'ils veulent dire est qu'une douleur est le genre de choses sur la présence duquel on peut à la rigueur mentir mais certainement pas se tromper. Les théories qu'évoque Dennett considèrent que, puisque le mot «connaissance» ne peut être utilisé de façon significative que là où des choses comme la simple supposition, le doute et l'erreur sont possibles en principe, nous ne devrions pas parler de connaissance là où elles sont exclues a priori. Un énoncé comme «J'ai mal» n'exprime pas une connaissance qui a l'avantage d'être plus assurée que n'importe quelle autre, il n'exprime pas véritablement une connaissance. Ce n'est pas un avantage, mais au contraire une faiblesse rédhibitoire pour une connaissance supposée que d'être constitutivement infaillible.

La douleur et son expression

La difficulté évidente de toutes les conceptions de ce genre est que ce qui est peut-être vrai pour des énonciations qui sont encore très voisines du simple cri ou gémissement de douleur, et ne sont effectivement pas des descriptions vraies ou fausses de quoi que ce soit, cesse de l'être pour des déclarations qui ont un contenu descriptif et informatif évident. Le fait qu'elles soient incorrigibles et infaillibles, si elles le sont réellement, n'autorise manifestement pas à nier qu'elles fassent référence à un fait déterminé et le décrivent de façon plus ou moins précise. Comme l'écrit Dennett: «Lorsque je dis au docteur que la douleur est dans le gros orteil, je ne fais certainement pas que produire une forme sophistiquée de gémissement, comme le suggère la conception de Wittgenstein, car j'ai tout à fait l'intention d'*informer* le docteur». La conception de Ryle souffre d'une insuffisance parallèle, et les deux conceptions, aussi plausibles que l'on puisse les rendre pour les comptes rendus de douleur, deviennent hautement implausibles lorsque d'autres énonciations introspectives sont prises en considération. La conception d'Anscombe est plausible jusqu'au moment où l'on demande comment elle propose de distinguer le fait que je *peux dire* toutes sortes d'absurdités du fait que je *peux dire* où est ma douleur» (id. pp. 100-101). Lorsque j'ai mal, je ne suis pas simplement dans une situation qui fait que je peux dire «J'ai mal»; je peux aussi dans un bon nombre de cas dire où j'ai mal, de quel genre est la douleur, etc. Et la question de la vérité et celle de savoir comment je peux connaître la vérité réapparaissent forcément.

Dennett rapporte que Ryle lui a exposé une conception destinée à rendre plus acceptable ce qu'il avait écrit dans *The Concept of mind* et consistant à introduire une stratification. A une extrémité du spectre, celle des aveux, on trouve de simples exclamations comme «Aïe!», qui sont purement expressives et évidemment pas susceptibles d'être dites vraies ou fausses; à l'autre extrémité, celle des comptes rendus ou des rapports caractérisés, on trouve des déclarations comme: «La douleur est dans la troisième dent en haut à gauche». «C'est, remarque Dennett, plausible, mais de toute évidence exige soit une explication de la manière dont les comptes rendus vrais à une extrémité sont infaillibles soit la conception implausible selon laquelle seuls les aveux sont soustraits à l'erreur» (p. 100, note 2). En d'autres termes, on comprend très bien qu'une énonciation qui est de la nature d'un cri ou d'une exclamation ne soit pas sujette à l'erreur; mais les énonciations qui se trouvent à l'autre extrémité du spectre et qui ont un contenu descriptif indiscutable ne semblent pas l'être davantage; et leur infaillibilité ne peut évidemment pas être expliquée de la même façon. Bien entendu, la conception de Wittgenstein est loin d'être aussi simpliste que le suggère la présentation de Dennett. Wittgenstein propose d'admettre que l'enfant qui apprend à utiliser un mot comme «douleur» apprend à remplacer des expressions naturelles inarticulées de la sensation par des expressions artificielles articulées et élaborées. Il apprend une manière nouvelle de se comporter lorsqu'on a mal[4]. Ce comportement est précisément nouveau et certainement tout à fait irréductible à l'ancien qu'il remplace. La remarque selon laquelle «l'expression verbale de la douleur remplace les cris et ne les décrit pas» ne signifie certainement pas que l'expression verbale est dans tous les cas assimilable à un cri: elle peut en être relativement proche ou au contraire tout à fait éloignée. D'autre part, cette remarque constitue essentiellement une critique du behaviorisme, et non comme semble le croire Dennett, de la conception cartésienne. Wittgenstein veut dire qu'il est absurde de soutenir que le mot «douleur» désigne une forme de comportement, telle que par exemple le fait de se contorsionner ou de pousser des cris; et il donne comme raison le fait que l'expression verbale se substitue justement au comportement et ne le décrit pas. Celui qui dit «J'ai mal» ne le fait évidemment pas sur la base de l'observation de son propre comportement et ne décrit pas le comportement en question. Et lorsqu'on dit de quelqu'un d'autre qu'il a mal, on ne veut pas dire qu'il se conduit d'une certaine façon, mais bel et bien ce que l'on dit, à savoir qu'il a mal.

[4] *Ein neues Schmerzbenehmen, Recherches philosophiques*, § 244.

Le fait que l'expression linguistique plus ou moins élaborée de la douleur ne décrive pas le comportement ne l'empêche évidemment pas de décrire autre chose, à savoir précisément une sensation. Il est vrai que Wittgenstein soutient que, si ce qui justifie mon affirmation «J'ai mal» n'est pas l'observation de mon comportement, ce n'est pas non plus l'observation d'un état de chose interne dont je constate préalablement l'existence. Lorsque je dis «J'ai mal», je n'identifie pas d'abord par l'observation interne mon état comme étant de l'espèce qu'on appelle «douleur» pour informer ensuite les autres de ce qu'il est. C'est tout à fait évident lorsque «J'ai mal» est encore suffisamment proche du simple cri de douleur, qui n'implique manifestement aucune identification ni description de ce que j'ai. Mais les choses deviennent malheureusement beaucoup plus compliquées lorsque la sensation est réellement décrite. Wittgenstein ne conteste évidemment pas qu'une sensation puisse être décrite de bien des façons. Ce qu'il veut dire est simplement qu'il y a des espèces très différentes de descriptions et que ce que nous appelons la description d'une sensation est réellement une espèce très particulière de description: «Songez combien il y a de choses de nature différente que nous appelons 'descriptions': description de la position d'un corps par ses coordonnées; description d'une expression faciale; description d'une sensation tactile; d'un état d'âme»[5]. «La difficulté est que, même si 'j'ai mal aux dents' est une description, c'est une description qui par sa forme a l'inconvénient de ressembler beaucoup plus qu'il ne le faudrait à 'j'ai 5 shillings'»[6]. De même le mot «assertion» utilisé à propos de «J'ai mal aux dents» est trompeur, parce que, lorsqu'on parle d'assertion, on s'attend à rencontrer un certain nombre de caractéristiques qui sont justement absentes dans le cas précis: «Appeler l'expression d'une sensation une *assertion* nous induit en erreur en ceci qu'au mot 'assertion' sont rattachés dans le jeu de langage de l' 'examen', la 'justification', la 'confirmation', l' 'infirmation' de l'assertion»[7]. Bien loin de chercher à réduire tout ce que nous serions tentés d'appeler des descriptions d'états internes à de simples expressions comportementales de ces états qui ont avec eux le même genre de relation que le cri avec la douleur qui le fait pousser, Wittgenstein souligne au contraire qu'il devient rapidement difficile, pour ne pas dire impossible, de déterminer si l'on a encore affaire à une simple description verbale de l'état, qui est une partie ou un aspect du comportement de quelqu'un qui est dans l'état en question ou, au contraire, à une authentique description d'un état interne: «Si quelqu'un dit 'J'espère qu'il viendra' – est-ce une *relation* (Bericht) sur son

[5] *Recherches philosophiques*, § 24.

[6] «Notes on Private and Sense Data», *Philosophical Review*, 77 (1968), p. 302.

[7] *Zettel*, B. BLACKWELL, Oxford, 1967, § 549.

état psychique, ou une *expression* (Aeusserung) de son espoir?»[8]. Si quelqu'un dit «J'ai peur», est-ce un cri de peur plus ou moins instinctif ou une information communiquée intentionnellement sur ce qu'il ressent ou encore une réflexion à usage plus ou moins personnel sur son état actuel? Wittgenstein note qu'il n'est probablement pas toujours possible de donner une réponse claire, ce qui ne signifie pas qu'il n'est jamais possible d'en donner une (cf. PU, p. 187).

La position qu'il défend est finalement bien plus hésitante et plus nuancée que celle qu'on lui attribue généralement sur la base de ce qu'il suggère à propos de l'apprentissage du mot «douleur»:

«Il y a tout de même ce problème-ci: le cri, que l'on ne peut pas appeler une description, qui est plus primitif que n'importe quelle description, n'en remplit pas moins la fonction d'une description de la vie psychique.

Un cri n'est pas une description. Mais il y a des transitions.

Et les mots 'J'ai peur' peuvent être plus ou moins proches et plus ou moins éloignés d'un cri. Ils peuvent s'en rapprocher de très près et ils peuvent en être éloignés d'une *très* grande distance.

Nous ne disons tout de même pas sans restriction de quelqu'un qu'il se *plaint,* parce qu'il dit qu'il a mal. Par conséquent, les mots 'J'ai mal' peuvent être une plainte, et peuvent être aussi autre chose.

Mais si 'J'ai peur' n'est pas toujours, et est cependant parfois, quelque chose qui ressemble à la plainte, pourquoi doit-il être *toujours* une description d'un état psychique?» (PU, § 1899).

L'infaillibilité des descriptions d'états internes

Il va de soi que ces considérations ne suffisent pas à résoudre le problème de Dennett, à savoir la question: qu'est-ce qui, en l'absence d'une faculté introspective infaillible, peut bien conférer aux descriptions d'états internes à la première personne la certitude absolue qu'on leur reconnaît? Dennett estime que les trois conceptions qu'il évoque sont sur la bonne voie lorsqu'elles essaient d'éviter le recours à la conception cartésienne du rapporteur infaillible. L'idée du rapporteur infaillible résulte en effet d'une transposition analogique qui implique que, dans le cas précis, ou bien il n'y a pas de rapporteur ou bien il n'est pas infaillible: «Puisqu'un rapporteur, un être humain peut commettre une erreur d'identification sur ce qu'il voit (sur ce que sont les choses là devant lui), le faire passer simplement 'à l'intérieur' et le transformer en une espèce quelconque d'introspectant n'est pas de nature à assurer qu'il rapportera infailliblement des expériences (ce que

[8] *Recherches philosophiques,* § 585.

sont les choses là à l'intérieur). On ne peut pas avoir des rapports sans un rapporteur, donc la notion de rapporteur infaillible doit être tout simplement erronée» (*op. cit.*, p. 101).

De façon générale, lorsqu'on considère la conscience comme une perception interne, ce qu'ont fait la plupart des théories traditionnelles, il est difficile de comprendre pourquoi de genre de perception n'est pas sujet à des erreurs et à des illusions du genre de celles qui affectent la perception externe, et cela d'autant plus que les théories actuelles de la perception ont tendance à insister sur ce qu'il y a d'hypothétique et de faillible dans toute perception. Selon Dennett, les conceptions de Wittgenstein, de Ryle et d'Anscombe empruntent la mauvaise voie lorsqu'elles cherchent à obtenir la solution à un niveau auquel elle n'est pas possible, à savoir le niveau personnel de l'explication: «Les trois conceptions nient toutes, du point de vue du discours que l'on tient dans le langage ordinaire sur les douleurs, les pensées, et ainsi de suite, que les énonciations introspectives soient – de ce point de vue – ce qu'elles sont de façon si évidente: des *rapports* de douleurs, de pensées, et ainsi de suite, qui peuvent, comme n'importe quels rapports, être vrais ou faux. Le rapporteur d'expériences mentales est, comme chacun sait, la *personne* elle-même, et ce qu'il fait consiste à rapporter, non à gémir ou à avouer ou à s'engager dans une sorte de glossolalie à laquelle les questions de vérité ne s'appliquent pas. Nous ne pouvons pas répondre à la question de savoir comment ces rapports sont infaillibles en niant qu'ils soient des rapports» (p. 101). Comme on l'a vu, Wittgenstein se garde bien de soutenir que la personne qui apparemment rapporte ses expériences mentales fait en réalité dans tous les cas quelque chose de différent et qui n'a rien à voir avec ce qu'il donne l'impression de faire. Ce qu'il dit est simplement qu'un énoncé comme «J'ai peur» ou «J'ai mal», dont la forme est celle d'un rapport ou d'une description, peut avoir dans certains cas une fonction bien différente et qu'avant de pouvoir être rapportée ou décrite, une expérience mentale comme celle de la peur ou de la douleur doit d'abord pouvoir être exprimée ou extériorisée sous des formes qui n'ont rien de descriptif. Il n'en reste pas moins qu'aucune réponse satisfaisante n'a été apportée à la question de savoir à quoi est due l'infaillibilité des rapports introspectifs, lorsqu'ils sont bien ce qu'ils ont l'air d'être, et non pas des choses à propos desquelles la question de la vérité ou de la fausseté ne se pose pas. On ne peut pas se borner à enregistrer simplement le fait qu'ils sont infaillibles. Mais, selon Dennett, c'est la seule chose que nous puissions faire, si nous en restons au niveau personnel, celui des relations entre une personne et ses expériences mentales: «Si nous ne sommes pas satisfaits – comme je pense que nous ne pouvons pas l'être – de voir l'explication arrêtée prématurément ici, à savoir de l'idée que les rapports introspectifs sont tout bonnement infaillibles, nous devons abandonner le niveau personnel et poser une question

différente: comment les énonciations introspectives peuvent-elles être reliées à certaines conditions internes d'une manière telle qu'elles peuvent être considérées comme des indications non sujettes à l'erreur de ces conditions internes?» (p. 101).

Au niveau subpersonnel, l'élément crucial pour la solution du problème réside dans la distinction entre un état fonctionnel ou logique d'un système et un état physique de ce même système. Putnam a été le premier à attirer l'attention sur l'analogie qui existe «entre les états logiques d'une machine de Turing et les états mentaux d'un être humain, d'une part, les états structuraux d'une machine de Turing et les états physiques d'un être humain, d'autre part»[9].

Une machine peut-elle simuler l'introspection?

Une machine de Turing peut être décrite logiquement, c'est-à-dire indépendamment de toute référence à son mode de réalisation physique, à l'aide d'une collection ordonnée d'états logiques ou fonctionnels qui sont complètement spécifiés dans la table de la machine par les relations qu'ils entretiennent les uns avec les autres et avec les entrées et les sorties de la machine. La machine de Turing, en tant qu'entité abstraite, est identifiée par sa table, c'est-à-dire par l'interrelation fonctionnelle de ses états, et non par sa constitution physique, qui n'est pas précisée et qui pourra être aussi différente qu'on voudra; et de la même façon un état logique de la machine est l'état qu'il est uniquement en vertu de ses relations à d'autres états et à l'entrée et à la sortie, et non de son mode de réalisation et de ses caractéristiques physiques. Une machine particulière T est dans l'état logique A si et seulement si elle effectue ce que la table de la machine spécifie pour l'état logique A, quel que soit l'état physique dans lequel elle se trouve au moment considéré. En d'autres termes, la nature exacte des états successifs par lesquels passe la machine qui effectue un calcul est indifférente, aussi longtemps que ces états sont distincts et se succèdent de la façon qui est spécifiée par la table de la machine, c'est-à-dire par les relations qu'ils doivent avoir entre eux et avec ce qui apparaît sur le ruban. Bien entendu, la machine qui pour effectuer un certain calcul, par exemple pour déterminer la 3000^{e} décimale de π, passe par une succession d'états A, B, C, etc. n'a pas besoin de reconnaître en un sens quelconque qu'elle passe par la suite d'états en question: il suffit qu'elle le fasse effectivement. D'autre part, si la machine se conduit ou opère comme si elle était dans l'état B, alors elle est par définition dans l'état B. Il est possible que, par suite d'un

[9] «Minds and Machines», in A. R. ANDERSON *(ed.), Minds and Machines,* Prentice-Hall, Inc. Englewood Cliffs, N.J. 1984, p. 84.

accident quelconque, elle se trouve dans l'état B, au moment où elle devrait se trouver dans l'état A. Mais si elle fait ce qu'elle ferait dans l'état B, alors elle est incontestablement dans l'état B.

Supposons à présent que la table de la machine contienne l'instruction: «Imprimez: 'Je suis dans l'état A' lorsque vous êtes dans l'état A». Lorsque la machine imprime «Je suis dans l'état A», devons-nous dire qu'elle a reconnu qu'elle était dans l'état A? D'un côté, il semble que pour appliquer l'instruction, la machine doive d'abord déterminer dans quel état elle est. Mais de l'autre, comme le fait remarquer Putnam, le «rapport verbal» que la machine produit sur l'état dans lequel elle se trouve «résulte directement de l'état qu'il 'rapporte': aucun 'calcul' ou 'élément de preuve' supplémentaire n'est requis pour parvenir à la réponse» (*ibid.*, p. 81). Le rapport résulte directement de l'état qu'il rapporte en ce sens que la machine est dans l'état A seulement si elle rapporte qu'elle est dans l'état A. Si elle imprimait qu'elle est dans un état différent, B, elle se comporterait ou opérerait comme si elle était dans l'état B et ne serait par conséquent pas dans l'état A. Si l'on peut donner un sens quelconque à la question «Comment la machine sait-elle qu'elle est dans l'état A?», la seule réponse possible est: «En étant dans l'état A». Il y a une certaine analogie entre cette situation et celle de quelqu'un qui a mal. Si l'on se demande comment il sait qu'il a mal, la réponse sera probablement qu'il le sait par le simple fait qu'il a mal. Il n'est pas possible d'avoir mal sans savoir en même temps qu'on a mal. La tendance de Wittgenstein est justement de soutenir qu'il n'y a au fond rien de plus dans «Je sais que j'ai mal» que dans «J'ai mal». Entre «J'ai mal» et «Je sais que j'ai mal», il n'y a pas de place pour un acte de connaissance qui viendrait s'ajouter à la simple présence de la sensation de douleur. On objectera naturellement que la machine passe successivement par différents états, mais qu'elle ne les «a» pas, au sens auquel on peut avoir une sensation de douleur et qu'elle en a encore moins une connaissance quelconque, même si elle est capable de produire l'énoncé «Je suis dans l'état A» toutes les fois qu'elle est dans cet état. Mais c'est oublier justement que nous raisonnons au niveau subpersonnel. Si la machine était le genre de chose à propos duquel nous avons de bonnes raisons d'adopter le niveau d'explication personnel, nous dirions probablement qu'elle a une connaissance infaillible de l'état dans lequel elle se trouve. Ce qui, au niveau d'explication personnel, est conçu comme résultant de l'exercice d'une faculté de connaissance infaillible est expliqué, au niveau d'explication subpersonnel, par l'impossibilité dans laquelle se trouve la machine de rapporter qu'elle est dans l'état A si elle n'est pas effectivement dans l'état A.

Même si un accident avait pour conséquence que la machine imprime à un moment donné «Je suis dans l'état A», alors qu'elle n'est pas dans l'état A, il ne s'agirait pas d'une erreur de calcul, mais plutôt de quelque chose

qui est de la nature d'un lapsus, tout comme quelqu'un qui affirmerait en toute sincérité qu'il a mal, alors qu'il éprouve une sensation qui n'a rien de douloureux, ne pourrait être soupçonné d'une erreur d'identification sur ce qu'il a réellement, mais plutôt d'une erreur ou d'une confusion verbales. Putnam contraste les énoncés «J'ai mal» (proféré par un être humain) et «Je suis dans l'état A» (imprimé par la machine) avec «J'ai de la fièvre» ou «Le tube à vide 312 n'a pas fonctionné». Les êtres humains disposent d'une certaine capacité d'information sur des états physiques internes; et de la même façon les ordinateurs peuvent être munis de dispositifs qui les tiennent au courant de leurs états physiques internes. Mais, dans les deux cas, la question de savoir comment les états en question ont été déterminés ou reconnus, a un sens clair. La réponse consiste à indiquer une succession d'états par lesquels le système doit passer pour reconnaître l'état physique dans lequel il se trouve. En revanche, si l'état rapporté est un état logique ou fonctionnel, la question de savoir comment le système le reconnaît, en est averti ou l'examine n'a pas de place dans le processus consistant à rapporter ce qu'il en est. Si l'on demandait à la machine comment elle sait que la 3000^{e} décimale de π est ce qu'elle dit qu'elle est, elle pourrait répondre en repassant par la suite des états qui ont abouti à la production du résultat, c'est-à-dire en réeffectuant le calcul. Reconnaître ou déterminer quel est le résultat veut dire appliquer la procédure qui conduit au résultat. Mais si une machine est capable en un sens quelconque de reconnaître dans quel état logique elle est, cela n'a pas de sens de se demander par quelle procédure elle aboutit à ce genre de résultat.

Celui-ci ne peut résulter que du fait qu'elle est effectivement dans l'état en question et non de l'utilisation d'une procédure de calcul. Calculer un résultat veut dire pour la machine passer par une certaine succession d'états logiques. Mais reconnaître l'état dans lequel elle est ne peut pas vouloir dire le calculer, c'est-à-dire passer par une succession d'autres états qui aboutit à la reconnaissance.

Dennett en conclut qu' «une machine de Turing conçue d'une manière telle que son *output* pourrait être interprété comme consistant en rapports de ses états logiques serait comme les introspecteurs humains, invulnérable à toutes les erreurs autres que 'verbales'. Elle ne pourrait pas *commettre une erreur d'identification* sur ses états logiques tout simplement parce qu'elle n'a pas du tout à identifier ses états» (*op. cit.* pp. 103-104). Cependant, pour rendre l'analogie avec l'introspection humaine vraiment intéressante, il est nécessaire de donner une idée beaucoup plus précise de ce que pourrait être une machine qui produit des comptes rendus introspectifs. Il serait évidemment tout à fait naïf de croire qu'une énonciation introspective ou d'ailleurs une production linguistique, quelle qu'elle soit, peut être une fonction directe d'états logiques intéressants quelconques, comme dans le cas de la machine qui imprime la phrase «Je suis dans l'état A» toutes les

fois qu'elle est dans cet état. La production du discours est médiatisée par des systèmes extrêmement complexes sur lesquels nous ne savons pour l'instant que très peu de choses, mais tels que nous sommes en mesure, depuis les travaux de linguistes comme Chomsky, de formuler un certain nombre d'hypothèses de nature générale sur ce qu'ils doivent être à partir d'un examen de la structure du langage lui-même. La situation n'est pas forcément très différente aujourd'hui de ce qu'elle était en 1969 au moment où Dennett a publié la première édition de son livre. Dennett raisonne à partir d'une situation hypothétique dans laquelle il serait devenu possible et souhaitable de construire une «machine percevante», c'est-à-dire une machine qui produit l'équivalent de jugements de perception tels que par exemple «Je vois un homme qui approche». Une telle machine serait dotée d'organes des sens qui pourraient être des caméras de télévision. L'information sensorielle qu'ils sont en mesure de recevoir sur le monde extérieur serait transmises à un ordinateur chargé de l'analyser. Et les sorties de l'analyseur serviraient d'entrées à un «centre de la parole» qui serait programmé de façon à les transformer en rapports sur ce qui est vu, c'est-à-dire en descriptions de la scène observée. Bien entendu, on est encore loin de pouvoir construire des systèmes qui disposeraient des compétences exigées de l'analyseur et du centre de la parole. Et, en outre, comme le remarque Dennett: «La machine percevante qui résulterait de toute cette compétence miraculeuse serait, bien entendu, une pâle copie d'un percepteur humain, puisque rien ne serait prévu qui lui permette d'utiliser ses 'perceptions' pour n'importe quelle autre fin que comme base de rapports verbaux, et la machine ne serait pas non plus dotée de la capacité de mentir sur sa vision des choses, de décider de parler d'un autre objet quelconque, de poser des questions, etc. Elle débiterait en toute naïveté des rapports de ce qu'elle voit, donnant des réponses presque skinnériennes à ses stimuli visuels. Mais elle partagerait une des caractéristiques cruciales du percepteur humain: elle ne pourrait pas commettre d'erreur sur ses états 'mentaux'» (p. 109).

Bien évidemment une machine de ce genre serait sujette à des erreurs qui constituent l'équivalent d'illusions ou d'hallucinations. On pourrait la tromper à l'aide d'un postiche en mouvement qui produit sur ses récepteurs sensoriels les mêmes impressions qu'un homme qui s'approche ou en fournissant directement à l'analyseur les données sensorielles qui correspondent à la perception d'un homme qui s'approche, à un moment où il n'y a aucun homme qui s'approche. La machine pourrait également commettre des «erreurs verbales»; mais, si l'on fait abstraction de celles-ci, toutes les données qui seront correctement exprimées relativement aux règles de langage auront été programmées dans l'ordinateur. La machine ne peut pas commettre des erreurs d'identification sur les données qui lui sont fournies par l'analyseur. La raison en est que le centre de la parole n'examine ni

n'analyse en aucune façon ses entrées pour déterminer leurs qualités ou leurs ressemblances et leurs dissemblances par rapport à d'autres entrées, il ne fait que produire des phrases françaises qui sont des *expressions* de ces entrées. Ce qui fait d'une sortie de l'analyseur la sortie qu'elle est, exactement ce qu'elle produira dans le centre de parole si elle lui est fournie comme entrée, de sorte qu'une sortie est exactement ce que le centre de parole considère qu'elle est, indépendamment de ses qualités et de ses caractéristiques dans une réalisation physique quelconque. En d'autres termes, la machine peut commettre des erreurs d'identification sur ce qu'elle a réellement devant les «yeux»; mais elle ne peut se tromper sur ce qu'il lui *semble voir*. Si au lieu de dire «Je vois un homme qui approche», elle se contentait d'écrire à chaque fois «Il me semble que je vois...» ou «C'est exactement comme si je voyais...», elle serait déchargée de toute responsabilité concernant les entrées frauduleuses ou les erreurs de l'analyseur et infaillible, aux erreurs verbales près.

Dennett décrit la situation ainsi: «Quelle que soit la forme des mots, quelle que soit la suite de symboles imprimés, ce qui est imprimé sera une expression de la sortie de l'analyseur; la forme verbale est simplement utilisée comme indicateur du fait que les discordances entre la sortie et le monde extérieur ne doivent pas être prises en considération. On pourrait tout aussi bien laisser tomber les rapports de la forme «Je vois...» et attacher un petit signe à la machine, «Pas responsable des entrées frauduleuses ou des erreurs dans l'analyse des entrées! Transposé au cas de l'énonciation humaine, ce point devient: l'immunité contre l'erreur n'a rien à voir avec l'exécution d'une *action personnelle* quelconque. Une explication de l'intention d'un homme, ou de *ce qu'il croit être en train de faire,* ne joue aucun rôle dans l'explication de la certitude introspective; quelle que soit l'*intention* avec laquelle une énonciation est produite (considérée au niveau personnel), au niveau subpersonnel elle sera une expression de l'entrée fournie au centre de parole humain (qui reçoit son entrée de sources plus nombreuses que simplement les analyseurs perceptuels), et *en tant que telle* elle sera exempte d'erreur relativement à la scène extérieure. En fait, bien entendu, lorsque nous entendons que nos énonciations soient immunisées de cette façon, c'est-à-dire lorsque nous entendons que les autres les jugent selon cet éclairage, nous donnons à nos expressions la forme de l'idiome «Il me semble que je vois...». En utilisant cet idiome, une personne n'exprime pas intentionnellement l'entrée de son centre de parole, car il est fort probable qu'elle n'a pas la moindre notion d'entrée du centre de parole; ce qui explique que l'énonciation soit immunisée contre l'erreur n'est rien que la personne *fasse* – n'est pas une *action* personnelle, intentionnelle ou autre – mais ce qui se passe dans son cerveau» (p. 111). Naturellement, lorsqu'une personne produit une description verbale de ses états mentaux, il s'agit généralement d'un acte intentionnel destiné à renseigner autrui sur ce

qu'ils sont. Mais l'infaillibilité de la description n'a rien à voir avec une intention ou une action quelconque. L'idée que quelque chose doit être examiné et reconnu sans aucune possibilité d'erreur pour que les rapports introspectifs puissent posséder l'infaillibilité qu'on leur attribue a tout simplement disparu de l'histoire.

L'introspection est-elle réellement infaillible?

Dans *Brainstroms* dont la première édition date de 1978, Dennett a modifié sur un certain nombre de points importants la conception qu'il avait esquissée dans *Content and Consciousness*. Mais la version plus développée et plus subtile qu'il propose conserve l'idée essentielle, qui est que l'infaillibilité au moins apparente (et peut-être, en fait, plus apparente que réelle) que possèdent les comptes rendus introspectifs doit pouvoir être expliquée dans les termes d'une connexion établie au niveau subpersonnel entre deux systèmes, sans qu'il soit nécessaire de faire intervenir une faculté d'observation particulière. L'accès privilégié de chacun à sa propre expérience est en fait déterminé par les relations d'accès qui existent entre les deux systèmes. Comme le dit Dennett: «Ainsi que l'exprimerait Anscombe il se trouve simplement que nous *pouvons dire* quelle chose nous sommes en train d'expérimenter, ce qu'il en est pour nous. Cela s'accomplit sans aucun œil interne ou faculté introspective en dehors de la machinerie invoquée dans le modèle»[10]. Dennett remarque que le problème a une certaine analogie avec celui que Hume pose à propos de la relation de cause à effet. Avant Hume, on avait tendance à considérer que lorsque nous voyons une cause et ensuite un effet, nous *voyons* la connexion nécessaire entre les deux, ce qui fait qu'ensuite nous inférons ou attendons la cause lorsque nous voyons l'effet. Hume montre que nous ne percevons aucune relation nécessaire de ce genre, mais tout au plus une relation de consécution constante. Les choses se passent en réalité dans l'ordre inverse de ce qu'on suppose généralement: nous avons été conditionnés à inférer ou à attendre l'effet lorsque nous voyons la cause et, nous trouvant en train d'effectuer l'inférence, nous sommes victimes de l'illusion que nous percevons une connexion nécessaire qui explique et justifie l'inférence que nous sommes irrésistiblement enclins à effectuer. Du point de vue psychologique et épistémique, c'est l'inférence qui est première et qui fait naître la croyance en une connexion nécessaire «perçue». C'est à peu près la même chose qui se passe dans le cas de l'introspection. «Je propose, écrit Dennett, une explication parallèle de l' 'introspection': nous nous trouvons en train de *vouloir dire* toutes ces choses sur ce qui se passe en nous; cela

[10] «Toward a Cognitive Theory of Consciousness», in *Brainstorms*, pp. 170-171.

donne naissance à des *théories* que nous soutenons à propos de la manière dont nous nous trouvons être capables de faire cela – par exemple, la théorie fâcheusement célèbre mais d'une parfaite simplicité selon laquelle nous 'percevons' ces choses qui se passent avec notre 'œil interne' et cette *perception* justifie et explique les intentions sémantiques que nous avons» (pp. 166-167). En d'autres termes, ce qui est fondamental et premier, ce sont les déclarations que nous sommes en mesure de produire et que nous éprouvons le besoin de produire sur toutes ces choses qui se passent en nous et qui, dans les conditions normales, sont considérées comme auto-certifiées et inaccessibles à une forme quelconque de contestation. L'introspection n'est pas un processus immédiatement appréhendable, elle est plutôt une construction destinée à expliquer quelque chose qui ne nous semble pas pouvoir être expliqué autrement. Si nous n'étions pas habitués à considérer et à exprimer les choses de cette façon, nous constaterions sans doute que même en «tournant la chose de tous les côtés», comme Hume le fait pour la cause, nous ne trouvons rien qui ressemble à ce dont nous sommes supposés avoir une perception immédiate.

Dennett explique: «Nos impressions de détenir une autorité spéciale lorsque nous formulons des rapports introspectifs – la base de toutes les thèses malformées de l'incorrigibilité et de l'infaillibilité introspectives – proviennent du fait que nos intentions sémantiques, qui déterminent ce que nous voulons dire, constituent les étalons par rapport auxquels nous mesurons nos propres productions verbales; par conséquent, si nous disons ce que nous avons l'intention de dire, si nous n'avons pas commis d'erreurs ou de fautes d'expressions, alors nos énoncés réels ne peuvent manquer d'être les expressions du contenu de nos intentions sémantiques, ne peuvent manquer de rendre justice à l'accès que nous avons à nos propres vies intérieures» (p. 171).

Observe-t-on ses états internes?

En fait, pour des raisons sur lesquelles il n'est pas possible de s'étendre ici, Dennett estime que c'était une erreur de parler, comme il l'a fait, d'infaillibilité ou d'incorrigibilité, d'avoir accepté trop facilement l'idée traditionnelle que «nous sommes les juges uniques et parfaits de ce que nous expérimentons» (p. 170). Au lieu d'incorrigibilité, il préfère parler simplement de ce que Gunderson appelle l' «asymétrie investigationnelle» de certaines déclarations à la première personne. Mais ce point est, de toute évidence, relativement secondaire par rapport à la question essentielle: l'idée que nous observons en quelque sorte nos états internes à l'aide d'un sens particulier a-t-elle réellement perdu toute espèce de respectabilité? Dans les *Recherches philosophiques,* Wittgenstein pose la question

suivante: «Celui qui observe son propre chagrin, avec quel sens l'observe-t-il? Avec un sens particulier. Avec un sens qui *sent* le chagrin: donc il le sent autrement lorsqu'il l'observe? Et lequel observe-t-il à présent: celui qui n'est là que lorsqu'il est observé?» (II, IX). Les théories traditionnelles supposent que l'on peut ressentir le chagrin et également observer le chagrin que l'on ressent. Mais si l'observer veut dire à nouveau le sentir et le sentir d'une certaine façon, le chagrin que l'on observe n'est pas le même que celui que l'on ressent et il n'existe que pendant qu'il est observé. C'est-à-dire que ce qu'on observe n'est pas l'état de choses indépendant de l'observation que l'on prétendait observer. Wittgenstein note que dans le cas normal ce qui est observé n'est pas le produit de l'observation. Mais si je dis par exemple: «Aujourd'hui je ne ressens plus la douleur que lorsque j'y pense», il y a bien quelque chose qui ressemble à une observation et une sensation qui est d'une certaine façon le produit de l'observation. Wittgenstein admet tout à fait qu'un énoncé comme «Mon chagrin n'est plus le même: un souvenir qui m'était encore insupportable il y a un an ne l'est plus pour moi aujourd'hui» est le résultat d'une observation. Mais justement les cas dans lesquels on peut parler réellement d'une attitude consistant à adopter la position de l'observateur et à essayer d'observer quelque chose en soi-même sont bien différents de ce qui se passe avec des énoncés comme «J'ai peur» ou «J'ai mal».

«Quand dit-on, se demande Wittgenstein, que quelqu'un observe? En gros: lorsqu'il se met dans une situation favorable pour recevoir certaines impressions pour (p. ex.) décrire ce qu'elles lui apprennent» *(ibid.)*. Il n'est guère contestable que, lorsque je dis «J'ai peur» ou «J'ai mal» je n'observe généralement rien en ce sens-là. Dans ses *Leçons sur la psychologie philosophique* de 1946-1947, Wittgenstein soulève la question générale suivante: «...Où est la science des phénomènes mentaux? Réponse: vous observez vos propres événements mentaux. Comment? Par l'introspection. Mais si vous observez, c'est-à-dire si vous vous mettez à observer vos propres événements mentaux, vous pouvez les modifier et en créer de nouveaux; et tout ce qui importe dans l'observation est que vous ne devriez pas faire cela, l'observation est supposée être précisément la chose qui évite cela. S'il en est ainsi, la science des phénomènes mentaux comporte le puzzle suivant: je ne peux pas observer les phénomènes mentaux des autres, et je ne peux pas observer les miens propres, au sens propre du mot 'observer' »[11].

[11] Lectures on *Philosophical Psychology*, 1946-47, ed. by P. T. GEACH, Harvester-Wheatsheaf, New York-London, 1988, p. 235.

Spinoza, dans un passage célèbre d'une lettre à Schuller, envisage le cas d'une pierre qui continue à se mouvoir après la cessation de l'impulsion externe, et fait la supposition suivante: «Concevez maintenant, si vous voulez bien, que la pierre, tandis qu'elle continue de se mouvoir, sache et pense qu'elle fait tout l'effort possible pour continuer de se mouvoir. Cette pierre, assurément, puisqu'elle n'est consciente que de son effort, et qu'elle n'est pas indifférente, croira être libre et ne persévérer dans son mouvement que par la seule raison qu'elle le désire. Telle est cette liberté humaine que tous les hommes se vantent d'avoir et qui consiste en cela seul que les hommes sont conscients de leurs désirs et ignorants des causes qui les font agir» (Ed. de la Pléiade, p. 1252). Pourrait-on dire la même chose de la faculté d'introspection que les hommes ont l'habitude de s'attribuer à l'égard de leurs états de conscience? Une machine qui aurait été construite et programmée de façon à ne pas pouvoir commettre d'erreurs autres que verbales sur ses états internes et qui ignorerait ce fait ne pourrait-elle pas s'imaginer qu'elle dispose d'une faculté d'observation interne qui lui permet de les reconnaître infailliblement? C'est au fond dans une situation de ce genre que nous pourrions nous trouver selon Dennett: la description que nous nous sentons obligés d'adopter au niveau d'explication personnelle pourrait n'être que le reflet de l'ignorance dans laquelle nous somme de certaines particularités de notre organisation au niveau subpersonnel.

L'objection qui vient immédiatement à l'esprit est qu'une machine de Turing a certes des états logiques, mais elle n'a pas d'états de conscience, pas de vie intérieure et pas la possibilité d'entretenir des illusions sur ce qu'elle fait. Nous nous heurtons ici à la difficulté générale que toutes les approches «fonctionnalistes» tentent avec plus ou moins de succès de surmonter: «Toute philosophie de l'esprit qui (comme moi-même) est en faveur d'une théorie 'fonctionnaliste' de l'esprit doit affronter le fait que la caractéristique précise dont on a vu qu'elle constituait une recommandation en faveur du fonctionnalisme par rapport à ce qui pourrait 'réaliser' les fonctions essentielles des systèmes sentants ou intentionnels – permet à une théorie fonctionnaliste, quel que soit le parfum de réalisme biologique ou humanoïde qu'elle dégage, d'être exemplifiée non seulement par des robots (une conséquence acceptable ou même désirable aux yeux de certains), mais par des organisations suprahumaines qui sembleraient avoir des esprits qui leur sont propres que dans le sens métaphorique le plus faible» (pp. 152-153).

Comme le dit Dennett: «Dans le meilleur des cas une théorie subpersonnelle semblera ne nous donner *aucune raison* de croire que ses exemplifications seraient des sujets doués d'une expérience et dans le pire (comme nous l'avons vu) une théorie subpersonnelle semblera permettre des exemplifications qui *ne sont manifestement pas* des sujets doués d'une expérience» (p. 154).

Il semblerait donc que toute théorie subpersonnelle doive nécessairement laisser de côté quelque chose d'essentiel. Mais, d'un autre côté, on ne peut pas non plus espérer expliquer ce qu'est une personne ou un moi si, pour ce faire, on doit utiliser d'entrée de jeu toutes les ressources descriptives qui ne sont accessibles qu'au niveau personnel. Le défi qu'ont relevé des gens comme Dennett est précisément celui qui consiste à essayer de constituer un «Je» à partir de parties subpersonnelles.

Ceux qui considèrent toute tentative de ce genre comme intrinsèquement déraisonnable en sont réduits, de leur côté, à constater simplement que nous n'avons pas affaire ici à un problème de réduction que l'on pourrait traiter par les moyens usuels, mais, au contraire, à une situation qui est, selon toute probabilité, condamnée à rester absolument singulière.

ERRATUM

Veuillez noter que deux membres de phrases manquaient dans l'article de J. Bouveresse, intitulé *L'«infaillibilité» de l'introspection* (vol. 122, 1990/II). Il faut lire (les mots en italiques manquaient):

p. 227, ligne 38: «La machine pourrait également commettre des «erreurs verbales»; mais, si l'on fait abstraction de celles-ci, toutes les données qui *sortent de l'analyseur pour entrer dans le centre de paroles* seront correctement exprimées relativement aux règles de langage *qui* auront été programmées dans l'ordinateur».

p. 232, ligne 27: «Toute philosophie de l'esprit qui (comme moi-même) est en faveur d'une théorie 'fonctionnaliste' de l'esprit doit affronter le fait que la caractéristique précise dont on a vu qu'elle constituait une recommandation en faveur du fonctionnalisme par rapport à *des variétés plus grossières de matérialisme – son caractère abstrait et partant sa neutralité par rapport à* ce qui pourrait 'réaliser' les fonctions essentielles des systèmes sentants ou intentionnels – permet à une théorie fonctionnaliste, quel que soit le parfum de réalisme biologique ou humanoïde qu'elle dégage, d'être exemplifiée non seulement par des robots (une conséquence acceptable ou même désirable aux yeux de certains), mais par des organisations suprahumaines qui sembleraient avoir des esprits qui leur sont propres que dans le sens métaphorique le plus faible» (pp. 152-153).

Avec toutes nos excuses.